INTERACTIONS BETWEEN THE AI CONTROL PROBLEM AND THE GOVERNANCE PROBLEM

Professor Nick Bostrom
University of Oxford
Director, Strategic Artificial Intelligence Research Center
Director, Future of Humanity Institute
Q: “If human level general AI is developed, then what are likely outcomes?”
<table>
<thead>
<tr>
<th>Date</th>
<th>Quote</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oct 2015</td>
<td>“Based on its level seen in the match (against Fan), I think I will win the game by a near landslide”</td>
</tr>
<tr>
<td>Feb 2016</td>
<td>“I have heard that Google DeepMind’s AI is surprisingly strong and getting stronger, but I am confident that I can win at least this time”</td>
</tr>
<tr>
<td>Mar 9, 2016</td>
<td>“I was very surprised because I didn’t think I would lose”</td>
</tr>
<tr>
<td>Mar 10, 2016</td>
<td>“I’m quite speechless ... I am in shock. I can admit that ... the third game is not going to be easy for me”</td>
</tr>
<tr>
<td>Mar 12, 2016</td>
<td>“I kind of felt powerless.”</td>
</tr>
</tbody>
</table>
Q:
“If human level general AI is developed, then what are likely outcomes?”
A:
Superintelligence
0 of 1 people found the following review helpful

⭐⭐⭐⭐⭐ One Star, September 10, 2016

By Amazon Customer

Verified Purchase (What's this?)

This review is from: Superintelligence: Paths, Dangers, Strategies (Paperback)

I read it in 3 days and I'm profoundly depressed.

Help other customers find the most helpful reviews

Was this review helpful to you? Yes No
Q: “What can we do now to maximize the probability of a positive outcome?”

A: • solve intelligence
  • solve scalable control
  • solve AI governance problem
technical research agendas

- inverse reinforcement learning
- adversarial examples
- models of control failure
- approval-maximizing agents
- imitation agents
- architectural composition
- corrigibility
- foundations of reflective agents
- detecting context change
- interpretability and explanation
- control diversification
the AI governance problem
Openness

- safety measures
- values
- (capability)
- source code, platforms
- science
- training data, environments, benchmarks

?
Observation

Openness reduces the gap between the leading developer and the nearest follower.

- a couple of years in a low openness scenario?
- a few months in a high openness scenario?
- zero in the limiting case of maximal openness

This could help reduce the risk that a small group monopolizes all the benefits.
Suppose that...

- safety requires some significant extra work after AI is completed

OR

- safe operation initially incurs a significant performance penalty

OR

- the Vulnerable World Hypothesis is true in the post-AI-transition world
Vulnerable world hypothesis

There is some level of technology at which offense strongly dominates defense, in the sense that any small group of reasonably competent people with access to the technology would be able to take some action that would lead to the destruction of the world (independently of what other people did after the action was taken).

biotechnology?
nanotechnology?
doomsday device?
Suppose that...

• safety requires some significant extra work after AI is completed

OR

• safe operation initially incurs a significant performance penalty

OR

• the Vulnerable World Hypothesis is true in the post-AI-transition world
Openness

- safety measures
- values
- (capability)
- source code, platforms
- science
- training data, environments, benchmarks
What to do?

• openness for now
• desired property: **conditional stabilization**…
• lay the foundations for a collaborative approach later:
  • coordinate (or ideally pool) research among trusted leading groups
  • create ability **not** to share science and algorithms until it is safe to do so
  • credibly commit to sharing benefits and influence